Controlling Social Media Data

Corone, Anna; Nanne, Annemarie; van Miltenburg, Emiel

Published in:
Proceedings of the 8th Conference on Computer-Mediated Communication CMC and Social Media Corpora (CMC-Corpora2021)

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Controlling Social Media Data: a Case Study of the Effect of Social Presence on Consumers' Engagement with Brand-generated Instagram Posts

Anna Corone, Annemarie J. Nanne, Emiel van Miltenburg
Tilburg University
E-mail: anna.corone@gmail.com, a.j.nanne@tilburguniversity.edu, c.w.j.vanmiltenburg@tilburguniversity.edu

Abstract

Research in social media marketing studies ways to increase customers' engagement with brand-generated social media posts. This can either be done through experiments, or corpus studies of existing social media posts. Experiments have the advantage that they are controlled, but they often lack ecological validity, while for corpus studies the reverse is often true. As a case study, we construct a corpus of 1761 brand-generated Instagram posts, looking at the effect of social presence (the perception of human contact) on different engagement metrics (likes and comments), taking the effect of possible confounds (theme of slogans, funniness, time) into account. We show how social media posts can be analyzed at different levels of granularity, to establish the strength of the effect of social presence. We hope that our work will help others to isolate the impact of different variables on post engagement on social media.

Keywords: social presence, customers' engagement, brand-generated social media posts

1. Introduction

Social presence (Short, Williams & Christie, 1976) is perceived when computer-mediated communication (CMC) tools are able to prompt the feeling of human contact through features enhancing warmth, personalization and sociability (Yoo & Alavi, 2001). Multiple studies focusing on social presence have shown that it is one of the main factors enhancing customers' engagement with social media posts (i.e., increasing the number of likes and comments) (Bakhshi, Shamma & Gilbert, 2014; Cyr et al., 2009). This is important because engagement is a sign of customers' satisfaction and emotional involvement with the brand, and therefore, it is the key to successful marketing strategies (Pansari & Kumar, 2017).

1.1 Related work

The presence of a face is an important feature that can influence social presence. The high number of neurons involved in the processing of a face resolves in increased attention to a stimulus when it contains a face compared to when it does not (Droulers & Adil, 2015). Therefore, the fact that faces are rich stimuli calls for a differentiation between levels of social presence intended as human cues in general and human faces in particular.

Earlier studies on social presence in computer-mediated-communication fall into two categories: controlled experiments and large-scale corpus studies. Cyr et al. (2009) provide an example of a controlled experiment. They distinguish three levels of social presence (low, medium, and high), and present users with carefully manipulated stimuli, to see if social presence affects user trust in e-commerce websites. An important contribution of their study is that while high levels of social presence (display of human faces) were the most appreciated, participants found medium levels of social presence (display of parts of the human body other than the face) confusing, but still better than the complete absence of a person. The downside of this study is that it is unclear to what extent their results generalize to real-world social media behavior. A way to examine real-world social media behavior is by a large-scale corpus study such has the one from Bakhshi, Shamma and Gilbert (2014). They automatically detect faces in a corpus of 1 million Instagram posts, and show that the presence or absence of faces is correlated with the number of likes and comments. While this study does provide an analysis of real-world social media behavior, because of the scale of their work, the data is relatively uncontrolled, and the authors are not able to use Cyr et al.'s more fine-grained distinction of low, medium, and high social presence. This paper explores the middle ground between the two approaches, where we select a specific set of posts, and carry out a controlled analysis of the individual factors influencing user engagement.

Figure 1: Overview of the Shanty Biscuits Instagram feed (@shantybiscuits), which is used for our case study.

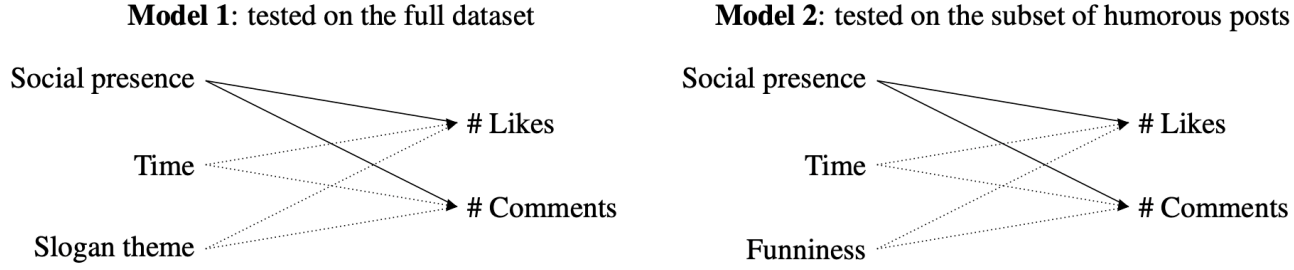

Figure 2: The two models used in our study. Both models use *social presence* as an independent variable, and *the number of likes/comments* on the post as a dependent variable. The models differ in their covariates (indicated with dotted lines). For Model 1, the covariates are *time since the post was placed*, and the *theme of the slogan*. For Model 2, covariates are *time since the post was placed*, and the *funniness of the post* (based on 10 ratings per post, see §2.2).

We present a case study, looking at the Instagram feed of Shanty Biscuits (@shantybiscuits), an existing French brand of personalized biscuits. Customers may submit slogans no longer than 12 characters including spacing or a logo to be printed on the biscuits. The company's Instagram feed is full of pictures of biscuits, apart from the text these posts differ relatively little from each other. There may or may not be a person holding the biscuit, and the backgrounds may be slightly different, but otherwise all the posts are very similar, as displayed in Figure 1. This allows for us to carry out a controlled study of the factors influencing user engagement, where our primary interest is the impact of low (no person visible), medium (part of a person visible), or high (face visible) social presence. Our study controls for the themes of the slogans, the funniness of the slogans, and the time when the picture was posted on Instagram. Based on previous studies, our **hypothesis** is that images with high social presence lead to more engagement (likes/comments) than images with low social presence and medium social presence (1a), which in turn leads to more engagement than pictures with low social presence (1b).

Thanks to the controlled setting of our study, we were able to isolate the effect of social presence from other confounds. This enabled us to examine real-world social media behavior while having a fine grained and controlled approach that helped us deal with the possible noise in the data. We found that the theme of the slogans can significantly affect whether social presence influences customers' engagement. This means that when analyzing the effect of social presence, it is important to consider also other potential factors that may confound its effect. We hope that our work contributes both to the study of social presence, and that our approach will help others aiming to isolate the impact of different variables on post engagement on social media.

2. Methods

With its people-centered focus, Instagram favors a higher degree of social presence (Bakhshi et al., 2014), making that social media platform suitable for the analysis here proposed. We carried out a content analysis of the entire Instagram feed of Shanty Biscuits. The metadata of the posts was retrieved through InstaLooter[1] on November 6, 2019. This data included the caption, the number of likes and comments, the date of publication, the Unix timestamp (the number of seconds elapsed since January 1, 1970 on the basis of which we calculated the number of days since the post was posted), the image URLs, the post URLs and the post ID of each published post. Subsequently, in order to delete possible duplicates among the posts, we used the Python Image Library[2] to find out which pictures had the exact same 200 pixels in a row, at a height of 200 pixels into the image. The present research proposes the analysis of the same hypothesis at two levels of granularity. For this purpose, two models as shown in Figure 2 are proposed. **Model 1** is tested on the whole dataset of the feed of the brand and includes social presence as independent variable, number of likes and comments as dependent variables, and the theme of the slogans of the cookies and time as covariates. **Model 2** is tested on the subset of humorous posts of the feed and includes social presence as independent variable, number of likes and comments as dependent variables and funniness perception of the posts and time as covariates.

2.1 Corpus study

The content analysis of a corpus of 2010 Instagram posts of Shanty Biscuits published between October 10, 2013 and September 30, 2019 was performed. After excluding 12 double posts with the Python Image Library, the remaining posts were manually coded in order to categorize them in three social presence categories and six text theme categories that were established throughout the analysis. A single annotator coded the Instagram posts. After the first round of coding, the same coder recoded the posts again in order to ensure reliability. At the end of the coding process, 231 posts were excluded as their pictures did not display the product and six posts were excluded as they were slideshows. At this stage, the final corpus consisted of 1761 brand-generated Instagram posts.

2.1.1. Social Presence

The three levels of social presence were coded consistent

[1] See: https://instalooter.readthedocs.io/en/latest/

[2] https://pillow.readthedocs.io/en/stable/

with the research of Cyr and colleagues (2009). The content analysis lead to the following categorization. **High social presence level** (*n*=146): included all the posts where the product appeared in the foreground and human facial features appeared entirely, partly or blurred in the background. **Medium social presence level** (*n*= 379): included posts in which parts of the human body other than the face were displayed. Most of the posts of this category show the product while being held by a hand. **Low social presence level** (*n*=1236): included posts where only the product appeared in a neutral background and no human images where displayed.

2.1.2. Theme of the slogans

With regards to the theme of the slogans, the analysis of the posts led to six categories. **Humorous text** (*n*=271): the slogans on the cookies of this category included funny quotes from movies, puns, sarcastic, silly, or mocking statements, and references to everyday problems. **Emotional events of life** (*n*=211): the slogans of the cookies of this category referred to the celebration of important events such as weddings, proposals, announcements of pregnancies, baby gender revelations, anniversaries and round-numbers birthdays. **Love** (*n*=191): this category included any text constituting a love declaration such as quotes from love songs or affectionate statements dedicated to partners, friends or family members. **No text** (*n*=149): this category included posts where the slogan of the cookies was not readable or cookies with logos rather than text. **Explicit content** (*n*=55): the slogan of the cookies of this category included explicit or sexual references. **Other** (*n*=884): is a final group which included posts in which the slogan on the cookies was not ascribable to any of the previous categories. Among others, it included external collaborations and partnerships, backstage moments of the company, special offers and discounts, and calls to action.

2.2. Online study

In order to analyze the posts at a finer level of granularity, we measured the funniness perception of the posts of the humorous subset and included the scores in the model as a covariate. For this purpose, an online study was conducted where people were asked to rate the funniness of a list of 30 or 31 posts randomly selected from the humorous subset. Before proceeding with the study, a manual check of the posts was performed, which led to the exclusion of 30 duplicates that were not detected by the Python Image Library. At this stage, the final subset of humorous slogans consisted of 241 Instagram posts. With a total of eight lists of posts and 80 participants, the online study guaranteed the collection of 10 funniness ratings per post. Since most of the slogans of the cookies were in French, participants were required to have at least a French proficiency at B1 level. This is recognized as the minimum level where non-native people are able to recognize different expressions of emotions (CEFR self-assessment grid) and have the necessary linguistic, pragmatic and sociolinguistic knowledge that is required to understand humor (Shively, 2013). Funniness perception was measured based on a three items 5-point Likert scale ranging from 'strongly disagree' to 'strongly agree' that was constructed based on existing literature about humor. The items developed to measure humor appreciation were created on the basis of an existing item adapted from Binsted, Pain and Ritchie (1997) (i.e. 'I found it funny') and two items created on the basis of the study from Chapman and Chapman (1974) (i.e. 'It made me laugh', 'It made me smile') where they stated that laughter and smiling are signs of humor appreciation.

2.2.1. Humor

The humorous subset was chosen for three main reasons. First, it was the theme category with the highest number of posts. Second, it was the most homogeneous category. In fact, in comparison with categories such as 'emotional events of life' or 'love', the posts present a lower degree of variation, as they mainly differ for social presence level and actual text. The other two categories, instead, showed a high degree of variation by often displaying different elements in the background such as people of different age and race, and having texts related to very different events of life such as weddings, birthdays or proposals that could also confound people's engagement with the posts. Third, past research showed a positive relationship between the use of humor on brand-generated social media posts and their number of likes and comments (Malhotra, Malhotra & See, 2013; Lee, Hosanagar & Nair, 2018).

2.2.2. Procedure

Participants were recruited through snowball sampling via Facebook groups of university students majoring in French language or personal connections via social media such as Facebook, Instagram and LinkedIn, and Survey Circle. Participants accessed the study via a link or a QR-code. After giving their informed consent, they were asked some demographic questions. Additionally, non-native speakers of French were asked to self-evaluate their level of French by choosing which of the six statements each corresponding to a level of the CEFR described their level best. Next, participants rated their funniness perception of the list of posts they were assigned to. Finally, they had the option to leave a comment before reading the debriefing.

2.3. Time

Having to deal with real-world data, means having to deal with noise. The main source of noise in our experiment is that posts that are public for longer have a higher potential to collect likes and comments through time than more recent posts. Therefore, we calculated how many days were elapsed since the post had been published, and included this count as a covariate in our two models.

3. Statistical analysis and results

In order to test the proposed hypothesis first at a coarse-grained and then at a fine-grained level, the two models

presented were tested performing a two-way MANCOVA using IBM SPSS 24.

3.1. Model 1: general effect of social presence

Recall that this model includes social presence as independent variable, number of likes and comments as dependent variables, and the theme of the slogans of the cookies and time as covariates. The results of the MANCOVA show that the covariate time is significantly related to customer's engagement overall $F(2, 1741) = 810.38$, $p < .001$, Pillai's Trace = .482, partial $\eta^2 = .48$. Specifically, the covariate is significantly related both to the number of likes, $F(1,1742) = 51.82$, $p < .001$, partial $\eta^2 = .03$, and to the number of comments, $F(1,1742) = 1142.12$, $p < .001$, partial $\eta^2 = .40$. There was a statistically significant difference between the levels of social presence on the combined dependent variables after controlling for time, $F(4, 3484) = 3.29$, $p = .011$, Pillai's Trace = .008, partial $\eta^2 = .004$, showing that overall social presence has an effect on customers' engagement. However, there was no significant individual effect of social presence on number of likes after controlling for time, $F(2,1742) = 2.54$, $p = .079$ and no significant individual effect of social presence was found on the number of comments after controlling for time, $F(2,1742) = 0.14$, $p = .870$.

Interestingly, the results showed that there was a statistically significant interaction effect between social presence and theme of the text on the combined dependent variables after controlling for time, $F(20, 3484) = 1.91$, $p = .008$, Pillai's Trace = .022, partial $\eta^2 = .011$. Specifically, a significant interaction effect of social presence and theme of the text on the number of likes of the posts after controlling for time was found, $F(10, 1742) = 2.77$, $p = .002$, partial $\eta^2 = .016$. However, no significant interaction effect between social presence and theme of the text when controlling for time was found for the number of comments, $F(10, 1742) = 0.89$, $p = .546$, partial $\eta^2 = .005$. The simple effect analysis showed that there is a significant effect of social presence levels on the number of likes for the 'love' theme category, such that consistent to hypothesis 1a, high levels of social presence ($M = 1373.70$, $SD = 933.02$) lead to higher numbers of likes than medium ($M = 362.55$, $SD = 501.42$), $p = .016$ and low levels of social presence ($M = 237.40$, $SD = 481.26$), $p = .017$. Moreover, an effect consistent to hypothesis 1b and partially consistent to hypothesis1a was found for the 'other' theme category, such that high levels of social presence ($M = 1202.58$, $SD = 1047.27$) lead to higher numbers of likes than low levels of social presence ($M = 417.33$, $SD = 606.11$), $p = .004$, and medium levels of social presence ($M = 1035.16$, $SD = 1418.89$) lead to higher numbers of likes than low levels of social presence, $p < .001$. At a coarse grained level, the results provided partial support for the expected hypothesis, showing that the expectation that higher levels of social presence lead to higher customers' engagement compared to lower levels of social presence is true in terms of number of likes for some themes of the slogans. Consequently, present results suggest that social presence does not have an effect on customers' engagement alone, but is also dependent on the type of posts it is related to, calling for a more fine-grained approach which focuses on a specific category of slogans.

3.2. Model 2: social presence in humorous posts

Recall that this model includes social presence as independent variable, number of likes and comments as dependent variables and funniness perception of the posts and time as covariates. The results of the MANCOVA show that the covariate time is significantly related to customer's engagement overall $F(2, 235) = 82.78$, $p < .001$, Pillai's Trace = .413, partial $\eta^2 = .41$. Specifically, the covariate is significantly related to both the number of likes, $F(1, 236) = 163.13$, $p < .001$, partial $\eta^2 = .41$, and comments, $F(1,236) = 25.12$, $p < .001$, partial $\eta^2 = .10$. The covariate funniness perception is not significantly related to customers' engagement overall, $F(2, 235) = 2.65$, $p = .073$, Pillai's Trace = .022, partial $\eta^2 = .022$. However, it is significantly related to the number of likes, $F(1, 236) = 4.98$, $p = .027$, partial $\eta^2 = .02$, but not to the number of comments, $F(1, 236) = 0.42$, $p = .520$, partial $\eta^2 = .002$. Moreover, there was no statistically significant difference between the levels of social presence on the combined dependent variables after controlling for time and funniness perception, $F(4, 472) = 0.58$, $p = .675$, Pillai's Trace = .010, partial $\eta^2 = .005$. Social presence does not have an effect on customers' engagement with humorous posts when controlling for both time and funniness perception, thus at a fine-grained level, the results show no support for the expected hypothesis.

4. Discussion

Being based on real-world data, the analysis posed some methodological challenges related to the control of noise in data. A lot of caution has been paid to create a corpus that was as much controlled as possible. The choice of the brand was a first step in that direction. In fact, Shanty Biscuits is a relatively small business that produces cookies which visually vary only in relation to the text impressed on them, resulting in a fairly homogeneous Instagram feed. However, the setting of the posts showed also some degree of variation in the levels of social presence and in the slogans impressed on the cookies. Therefore, the posts were categorized on the basis of these two factors in order to isolate those variables. Moreover, in order to control for possible confounds at different levels of granularity, two models and two approaches were proposed.

4.1. Model 1

The analysis of model 1 showed that the theme of the slogans significantly impacts the effect of social presence on customers' engagement intended as number of likes, but not as number of comments, such that for two particular themes social presence has indeed an effect on social presence consistent to our proposed hypothesis, whereas for the other themes no significant effect was found. The

results of the simple effect analysis show that the effect of social presence on the number of likes is significant for the posts of the category 'love' and 'others'. More specifically, for the category *Love* it was found that high levels of social presence significantly lead to higher numbers of likes than medium and low levels of social presence. For the category *Other*, the results show that high levels of social presence lead to a higher number of likes than low levels and that medium levels lead to higher numbers of likes than low levels of social presence, showing in both cases a partial support for the expected hypothesis. Why we only find an effect for a subset of the categories, is a topic for future research, but as we have seen in the results for Model 2, the content of the slogan (in the case of Model 2: funniness) is significantly related to the number of likes, so we expect the same to be the case here.

No significant effect of social presence was found on the number of comments under the posts, suggesting that particular attention to the possible difference between likes and comments should be taken into account in future research. Past research already pointed out that liking a post is quicker and less committing than commenting a post (Antheunis, van Kaam, Liebrecht & van Noort, 2016), which makes comments less direct. Moreover, commenters can also respond to each other, making the comment count a noisier signal to measure engagement. Future research could further investigate this difference to understand which other factors need to be taken into account when analyzing customers' likelihood to comment a post.

4.3. Model 2

The analysis of model 2 showed that no significant effect of social presence was found at this further level of granularity. Considering that the funniness perception was not a significant covariate in the model, the results of the analysis of the humorous subset are consistent with the ones of model 1, where no significant effect of social presence on engagement was found for the humorous slogans. The decision of focusing on the humorous subset for the fine-grained approach was taken before conducting the analysis of model 1, based on reasons that have been already explained. The results of the simple effect analysis in model 1 suggest that probably it would have been interesting to apply the fine-grained approach to the subset of love-related posts, in order to establish the strength of the effect of social presence at this finer level of granularity. Therefore, we suggest future research to take this aspect into account.

5. Conclusion

Present research proposed two possible approaches in order to test the effect of social presence on customers' engagement with brand-generated social media posts and to eventually establish its strength, while guaranteeing high ecological validity. Through a case study of an existing brand, the proposed hypothesis was tested at two levels of granularity. The aim of the research was to contribute to the question of how to control for variation when it comes to analyzing corpora of social media posts. Showing that indeed at different levels of granularity the strength of the effect of social presence changes, present research suggests that social presence needs to be analyzed in combination with other factors in order to have more accurate estimates of its effect.

6. References

Antheunis, M., van Kaam, J., Liebrecht, C., & van Noort, G. (2016). Contentmarketing op sociale netwerksites: Een onderzoek naar gedrag en motivaties van consumenten.*Tijdschrift voor Communicatiewetenschap*, *44*(4), pp. 337-365.

Bakhshi, S., Shamma, D. A., & Gilbert, E. (2014). Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 965-974.

Binsted, K., Pain, H., & Ritchie, G. D. (1997). Children's evaluation of computer-generated punning riddles. *Pragmatics & Cognition*, *5*(2), pp. 305-354.

Chapman, A. J., & Chapman, W. A. (1974). Responsiveness to humor: Its dependency upon a companion's humorous smiling and laughter. The *Journal of Psychology*, *88*(2), pp. 245-252.

Council of Europe. (n.d.). Self-assessment grid - Table 2 (CEFR 3.3): Common Reference levels. Retrieved from https://www.coe.int/en/web/common-european-framework-referencelanguages/table-2-cefr-3.3-common-reference-levels-self-assessment-grid

Cyr, D., Head, M., Larios, H., & Pan, B. (2009). Exploring human images in website design: a multi-method approach. *MIS quarterly*, *3*(33), pp. 539-566.

Droulers, O., & Adil, S. (2015). Could face presence in print ads influence memorization? *Journal of Applied Business Research (JABR)*, *31*(4), pp. 1403-1408.

Lee, D., Hosanagar, K., & Nair, H. S. (2018). Advertising content and consumer engagement on social media: evidence from Facebook. *Management Science*, *64*(11), pp. 5105-5131.

Malhotra, A., Malhotra, C. K., & See, A. (2013). How to create brand engagement on Facebook. *MIT Sloan Management Review*, *54*(2), pp. 18-20.

Pansari, A., & Kumar, V. (2017). Customer engagement: The construct, antecedents, and consequences. *Journal of the Academy of Marketing Science*, *45*(3), pp. 294—311.

Shively, R. L. (2013). Learning to be funny in Spanish during study abroad: L2 humor development. *The Modern Language Journal*, *97*(4), pp. 930-946.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. John Wiley & Sons.

Yoo, Y., & Alavi, M. (2001). Media and group cohesion: Relative influences on social presence, task participation, and group consensus. *MIS quarterly*, *25*(3), pp. 371—390.